10-708: Probabilistic Graphical Models 10-708, Spring 2014

16 : Markov Chain Monte Carlo (MCMC)

Lecturer: Matthew Gormley *Scribes: Yining Wang, Renato Negrinho*

1 Sampling from low-dimensional distributions

1.1 Direct method

Consider the one-dimensional case $x \in \mathbb{R}$ and that we would like to sample x from a complex distribution P. Let $h : \mathbb{R} \to [0, 1]$ be the cumulative density function (CDF) of P and suppose the *inverse* function h^{-1} is known. We can then sample x from P via the following procedure:

1. Sample u from $U[0, 1]$, the uniform distribution over $[0, 1]$.
2. Output $x = h^{-1}(u)$.

It can be easily proved that $x \sim P$ because for any $t \in \mathbb{R}$ we have that

$$\begin{aligned}\Pr[x \leq t] &= \Pr[h^{-1}(u) \leq t]\\ &= \Pr[u \leq h(t)] = h(t).\end{aligned}$$

This method is exact and is highly efficient when h^{-1} can be easily computed. However, when h is very complicated its inverse might not admit an easily computable form. In addition, the method is difficult to generalize to high dimensional cases when x has more than one covariate.

1.2 Rejection sampling

Suppose we want to sample x from P. Let Q be a distribution that is easy to sample from (e.g., uniform or Gaussian distribution) and let k be a constant such that $kQ(x) \geq P(x)$ for all x. The following procedure then produces a sample x that is exactly sampled from P:

1. Sample y from Q.
2. Sample u from $U[0, kQ(y)]$, where $U[0, kQ(y)]$ is the uniform distribution over interval $[0, kQ(y)]$.
3. If $u > P(y)$, discard y and repeat from the first step; otherwise, return $x = y$ as the sample.

The rejection sampling justified by the *envelope principle*: (y, u) are jointly sampled from the uniform distribution over the subgraph of $kQ(x)$. Thus accepting pairs with $u < P(y)/kQ(y)$ produces samples uniformly sampled from the subgraph of $P(x)$. As a result, the marginal distribution of y in the accepted pairs is exactly the same of P.

Rejection sampling is also exact and does not need to invert the CDF of P, which might be too difficult to evaluate. However, the rejection sampling procedure might reject a large number of samples before finally producing one, when the *envelope* distribution Q does not align well with the target distribution P. This disadvantage is even more serious in high-dimensional settings, thanks to the notorious *curse of dimensionality.*

1.3 Importance sampling

Unlike rejection sampling that aims at sampling from a particular distribution P, the importance sampling method concerns evaluating *statistics* under P, e.g., $\mathbb{E}_P[f(x)]$. Suppose Q is another distribution that is easy to sample with. The importance sampling procedure is as follows:

1. Draw S samples i.i.d. from distribution Q; denoted as $x^{(1)}, \cdots, x^{(S)}$.
2. Produce $\frac{1}{S}\sum_{s=1}^{S} f(x^{(s)})\frac{P(x^{(s)})}{Q(x^{(s)})}$ as an estimate of $\mathbb{E}_P[f(x)]$.

It is not difficult to prove that the estimate is unbiased; more specifically;

$$\begin{aligned}\mathbb{E}\left[\frac{1}{S}\sum_{s=1}^{S} f(x^{(s)})\frac{P(x^{(s)})}{Q(x^{(s)})}\right] &= \mathbb{E}_q\left[f(x)\frac{P(x)}{Q(x)}\right]\\ &= \int_{\mathcal{X}} Q(x)\cdot f(x)\frac{P(x)}{Q(x)}\mathrm{d}x\\ &= \int_{\mathcal{X}} f(x)P(x)\mathrm{d}x\\ &= \mathbb{E}_P[f(x)].\end{aligned}$$

Furthermore, the *variance* of the estimate decreases as we increase the number of samples S, rendering the resulting estimate more accurate. The estimation accuracy also depends on the alignment of the *proposal* distribution Q compared with the true distribution P to be evaluated.

1.4 Curse of dimensionality

Rejection/importance sampling usually behaves very poorly with increased dimensionality of $\mathcal{X}$. Consider, that $\mathcal{X} = \mathbb{R}^D$ and the true and the proposal (envelope) distributions defined as

$$P = \mathcal{N}_D(0, I), \qquad Q = \mathcal{N}_D(0, \sigma^2 I).$$

The rejection sampling must require $\sigma \geq 1$ to work. In this case, the probability of rejecting a proposed sample is σ^{-D}, which scales exponentially with the dimension D.

The importance sampling procedure requires $\sigma \leq 1/\sqrt{2}$ to have a well-defined finite variance of the resulting estimate. The variance (of $S = 1$) of the resulting estimator is

$$\left(\frac{\sigma^2}{2 - 1/\sigma^2}\right)^{D/2} - 1,$$

which again scales exponentially with the dimensionality D.

2 Markov Chain Monte Carlo

In this section we study MCMC methods to obtain a sequence of samples $\{x^{(t)}\}_{t=1}^{T}$ from an underlying distribution P. Suppose p is the pdf associated with P; we assume that

$$p = \frac{\tilde{p}}{Z},$$

where Z is a deterministic normalization constant that is difficult to compute and $\tilde{p}$ is a function of X that is easy to compute.

2.1 Metropolis Algorithm

In the Metropolis algorithm, we choose a simple *proposal* distribution $q(\cdot|x')$ that is easy to sample from and an initial sample $x^{(1)}$. Here q must be *symmetric*, meaning that

$$q(x|x') = q(x'|x), \qquad \forall x, x'.$$

We then perform the following steps for T times, each time obtaining a new sample $x^{(t+1)}$:

1. Propose a new sample $x \sim q(x|x^{(t)})$.
2. Compute the acceptance probability

$$a = \min\left(1, \frac{\tilde{p}(x)}{\tilde{p}(x^{(t)})}\right).$$

   Accept the new sample x with probability a.
3. If x is accepted, set $x^{(t+1)} = x$ and continue to the next iteration; otherwise repeat from the first step until x gets accepted.

2.2 Metropolis-Hastings Algorithm

In the Metropolis-Hastings algorithm, the proposal distribution q is no longer required to be symmetric. The sampling procedure is essentially the same as in the Metropolis algorithm, except the acceptance probability a' is computed differently as

$$a' = \min\left(1, \frac{\tilde{p}(x)q(x^{(t)}|x)}{\tilde{p}(x^{(t)})q(x|x^{(t)})}\right).$$

Needless to say, Metropolis algorithm is a special case of Metropolis-Hastings with symmetric proposal distributions q.

2.3 Gibbs sampling

Suppose x can be decomposed as $x = (x_1, \cdots, x_N)$. The Gibbs sampling procedure iteratively samples $x^{(t+1)}$ based on $x^{(t)}$ by performing the following steps:

1. Set $y^{(t+1)} = x^{(t)}$.

2. for all i in $\{1, \cdots, n\}$, sample $y_i^{(t+1)}$ from its conditional distribution $p(y_i^{(t+1)} | y_{-i}^{(t+1)})$.
3. Produce $x^{(t+1)} = y^{(t+1)}$.

It can be shown that Gibbs sampling is a special case of Metropolis-Hastings algorithm with proposal distribution $q(x_i|x_{-i}) = p(x_i|x_{-i})$. In particular, it can be shown that the acceptance probability a is always 1, as in the following derivation:

$$\begin{aligned}\frac{\tilde{p}(x)q(x^{(t)}|x)}{\tilde{p}(x^{(t)})q(x|x^{(t)})} &= \frac{\tilde{p}(x)p(x_i^{(t)}|x_{-i})}{\tilde{p}(x^{(t)})p(x_i|x_{-i}^{(t)})} \\ &= \frac{p(x)p(x_i^{(t)}|x_{-i})}{p(x^{(t)})p(x_i|x_{-i}^{(t)})} \\ &= \frac{p(x_i|x_{-i})p(x_{-i}) \cdot p(x_i^{(t)}|x_{-i})}{p(x_i^{(t)}|x_{-i}^{(t)})p(x_{-i}^{(t)}) \cdot p(x_i|x_{-i}^{(t)})} \\ &= 1.\end{aligned}$$

2.4 Markov Chain Monte Carlo in General

All three examples above are instantiations of MCMC algorithms. A MCMC algorithm involves the specification of a stochastic process defined through a Markov chain. If some conditions are satisfied, it can be shown that in the limit, after we run the Markov chain for a long time, we get an independent sample from the desired probability distribution p.

A Markov chain over some set of states $\mathcal{S}$, is specified by an initial probability distribution $p^{(0)} : \mathcal{S} \to \mathbb{R}_{\geq 0}$ over states and a transition probability operator $T : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$. The operator T determines the probability of transitioning from state x to state x'. It is assumed that the transition operator is the same across all time steps. The Markov chain is said to be homogeneous. Having some probability distribution over states $p^{(t)}$ at time-step t, the chain induces a probability distribution $p^{(t+1)}$ at time-step $t + 1$ by using the transition operator T as

$$p^{(t+1)}(x') = \int T(x'|x)p^{(t)}(x)dx.$$

To guarantee that the Markov chain does in fact define a process that samples from the desired probability distribution p in the limit, the following properties must be hold:

1. The desired distribution p must be invariant under the transitioning process of the Markov chain. It means that p is the stationary distribution of the chain. Formally,

$$p(x') = \int T(x'|x)p(x)dx.$$

2. The Markov chain must be ergodic. This means that irrespective of the choice of the initial probability distribution $p^{(0)}$, the chain converges to p:

$$p^{(t)}(x) \to p(x) \text{ as } t \to \infty, \text{for any } p^{(0)}.$$

Some of reasons under which the chain may fail to be ergodic are the existence of periodic (i.e., cyclic) behaviour, or the existence of two or more subsets in the state space that cannot be reached from each other. The convergence of the of distribution of states to the stationary distribution is usually referred to as mixing.

Detailed balance is a sufficient condition for the transition operator T under which the desired distribution p is an invariant of the chain. The condition is

$$T(x'|x)p(x) = T(x|x')p(x').$$

Detailed balance means that, for each pair of states x and x', to arrive at x and then transition to x' is equiprobable to arrive at x' and then transition to x. It is easy to verify that the condition implies the desired invariance. By integrating both sides, we get

$$\int T(x'|x)p(x)dx' = \int T(x|x')p(x')dx'.$$

We then notice that $\int T(x'|x)dx' = 1$, therefore

$$p(x) = \int T(x|x')p(x')dx',$$

which is the desired invariance condition. Furthermore, if p is invariant under the chain and

$$v = \min_{x,x';p(x')>0} \frac{T(x'|x)}{p(x)} > 0,$$

then p is the stationary distribution of the chain.

Going back to Metropolis-Hastings, it is easy to show that detailed balance is satisfied by its transition operator. The transition operator for Metropolis-Hastings is

$$\begin{aligned} T(x'|x) &= q(x'|x)a(x';x) \\ &= q(x'|x) \min\left(1, \frac{p(x')q(x|x')}{p(x)q(x'|x)}\right), \end{aligned}$$

where q is the proposal distribution and $a(x',x)$ is the probability of accepting state x', given that we are now in state x. We first propose a new state x' according to the proposal distribution q, and accept it with probability $a(x',x)$. We can now verify that detailed balance holds:

$$\begin{aligned} T(x'|x)p(x) &= p(x)q(x'|x) \min\left(1, \frac{p(x')q(x|x')}{p(x)q(x'|x)}\right) \\ &= \min\left(p(x)q(x'|x), p(x')q(x|x')\right) \\ &= p(x')q(x|x') \min\left(\frac{p(x)q(x'|x)}{p(x')q(x|x')}, 1\right) \\ &= p(x')T(x|x'). \end{aligned}$$

Another way of constructing a transition operator T that satisfies the invariance properties for p is by mixing or concatenating other transition operators $T_1, \ldots, T_n$ that also satisfy the invariance property. For mixture case,

$$T(x'|x) = \sum_{i=1}^{n} \alpha_i T_i(x'|x),$$

with $\sum_{i=1}^{n} \alpha = 1$ and $\alpha_i \geq 0$ for all $i \in [n]$, the invariance is easily shown:

$$\begin{aligned}
\int T(x'|x)p(x)dx &= \int \left(\sum_{i=1}^{n} \alpha_i T_i(x'|x) \right) p(x)dx \\
&= \sum_{i=1}^{n} \alpha_i \left(\int T_i(x'|x)p(x)dx \right) \\
&= \sum_{i=1}^{n} \alpha_i p(x) \\
&= p(x),
\end{aligned}$$

where we used the invariance of the individual $T_1, \ldots, T_n$ and $\sum_{i=1}^{n} \alpha_i = 1$.

2.5 Practical considerations of Markov Chain Monte Carlo

Markov chain Monte Carlo is widely applicable, but it is not without its problems.

While in theory, it is guaranteed to converge in the limit to the desired probability distribution p, in practice getting sufficiently close to convergence (i.e., mixing) may take a very large number of steps. Even worse is the fact that it is notoriously difficult to assess if the chain has converged or not. This means that answers computed using samples from a chain that has not mixed properly may be completely wrong in ways that are hard to diagnose.

MCMC methods usually have some number of hyperparameters, and the behaviour of the methods depends critically on the their value. One example is the scale parameter of transition operator T: if chosen too small, the process will move with very small steps through the state space; if chosen too large, most of the proposed steps will be rejected and the process will stay a long time in the same place; in both of these cases the chain will take a very long to converge. Setting the parameters incorrectly will lead to large mixing times, making the methods impractical. Even with appropriate settings for the parameters, mixing may still be very slow due to the random walk type of exploration of the space.

There is also a significant number of design decision and hyperparameters that are hard to set. These include what transition operator to use; what scale parameter for the transition operator; how many steps for the initial burn-in period; how many samples to collect; how to assess convergence; choosing to run one long chain versus multiple smaller chains; using all the samples of the chain for inference versus just a few independent ones, and if just a few, how many before we consider we have independent samples.

Some of these issues are addressed heuristically to some degree. For example, a common heuristic for the scale parameter of the transition operator T is setting it such that we accept half of the proposed steps. For assessing the convergence of the chain, it is common to run several chains in parallel with different initial conditions and gather statistics of the process. If the different chains look similar after some number of steps, we can be reasonably confident that the process has converged.

Nonetheless, these heuristics do not address the fundamental slowness of exploring the state space by means of a random walk. A rule of thumb is that if we need in the order of

$$\left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)^2,$$

to get independent samples from the chain, where $\sigma_{\max}$ and $\sigma_{\min}$ are the maximum and minimum length scales in the desired distribution p. Roughly speaking, the more correlated the random variables in the p are, the larger the mixing time of the chain.